Nature Conservation 35: 97–116 (2019)
doi: 10.3897/natureconservation.35.33918
http://natureconservation.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
Nature Conservation
Launched to accelerate biodiversity conservation

Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models

Rainer Ferdinand Wunderlich[1], Yu-Pin Lin[1], Johnathen Anthony[1], Joy R. Petway[1]

1 *Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.)*

Corresponding author: *Yu-Pin Lin* (yplin@ntu.edu.tw)

Academic editor: *Petr Keil* | Received 18 February 2019 | Accepted 15 May 2019 | Published 20 June 2019

http://zoobank.org/F39E11E0-2C30-4588-80B8-8B8415A9BD77

Citation: Wunderlich RF, Lin Y-P, Anthony J, Petway JR (2019) Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. Nature Conservation 35: 97–116. https://doi.org/10.3897/natureconservation.35.33918

Abstract

Model evaluation metrics play a critical role in the selection of adequate species distribution models for conservation and for any application of species distribution modelling (SDM) in general. The responses of these metrics to modelling conditions, however, are rarely taken into account. This leads to inadequate model selection, downstream analyses and uniformed decisions. To aid modellers in critically assessing modelling conditions when choosing and interpreting model evaluation metrics, we analysed the responses of the True Skill Statistic (TSS) under a variety of presence-background modelling conditions using purely theoretical scenarios. We then compared these responses with those of two evaluation metrics commonly applied in the field of meteorology which have potential for use in SDM: the Odds Ratio Skill Score (ORSS) and the Symmetric Extremal Dependence Index (SEDI). We demonstrate that (1) large cell number totals in the confusion matrix, which is strongly biased towards 'true' absences in presence-background SDM and (2) low prevalence both compromise model evaluation with TSS. This is since (1) TSS fails to differentiate useful from random models at extreme prevalence levels if the confusion matrix cell number total exceeds ~30,000 cells and (2) TSS converges to hit rate (sensitivity) when prevalence is lower than ~2.5%. We conclude that SEDI is optimal for most presence-background SDM initiatives. Further, ORSS may provide a better alternative if absence data are available or if equal error weighting is strictly required.

Keywords

Species distribution modelling, True Skill Statistic, evaluation, presence-background

Copyright *Rainer Ferdinand Wunderlich et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Species Distribution Modelling (SDM) relates independent environmental variables to species occurrence data and, in turn, predicts a dependent variable such as probability or the relative likelihood of occurrence (Guisan and Zimmermann 2000; Peterson 2001; Guillera-Arroita et al. 2015). Even though SDM predictions mostly range from zero to one, SDM predictions are often discretised into binary presence-absence maps (i.e. comprising only zeros and ones) used to evaluate wildlife management options, to identify appropriate conservation translocation sites and to evaluate model performance (Willis et al. 2009; Fordham et al. 2012; Liu et al. 2013) with confusion matrix-based performance metrics. These confusion matrices (Table 1) summarise the correspondence between predictions and observations, by providing the counts of (a) true presences, (b) false presences (commissions), (c) false absences (omissions) and (d) true absences. However, inherent asymmetric uncertainty levels, particularly for mobile species, between the observed and predicted presence and absence classes, can complicate such comparisons.

Observed absences in presence-absence datasets can be either true, i.e. the species does not occur, or false, i.e. the species does occur but remains undetected (Martin et al. 2005). 'Observed true absences' result from biological processes, such as intolerance to local conditions, competition (Hardin 1960; Leathwick and Austin 2001), general rarity (Gaston 1994), meta-population dynamics, i.e. perpetuating series of local extinctions and recolonisations (Hanski 1998) or other biotic interactions (Bascompte 2009; Wisz et al. 2013; Bulleri et al. 2016). 'Observed false absences', on the other hand, are artefactual in nature, resulting from insufficient monitoring relative to species movement (Tyre et al. 2003) or imperfect detection (MacKenzie et al. 2002). Whereas both true and false absences can lead to 'zero-inflated' datasets (Heilbron 1994) that violate statistical assumptions, the latter are also a source of uncertainty in parameter estimates as artefactual signals (e.g. sampling bias, probability of detection) confounding estimates of probability of occurrence (MacKenzie et al. 2002).

The capability to distinguish observed true and false absences may also dictate the applicability of model evaluation metrics, many of which differ in weights assigned to each of the four categories in the confusion matrix (Table 1). For example, when the observed true and false absences are indistinguishable, omission errors, i.e. excluding known presences (type c in Table 1), may be more problematic than commission errors, i.e. including absences of relatively unknown certainty (type b in Table 1). Although seldom investigated, differential error weighting may also be desirable when considering specific biological questions that relate to population dynamics, such as population

Table 1. Confusion matrix with cell designation as defined by the agreement of predictions (rows) and observations (columns).

Predicted	Observed	
	Present	***Absent***
Present	True (a)	False (b)
Absent	False (c)	True (d)

sources versus population sinks (Guisan and Thuiller 2005), or biogeographical processes like invasion source versus colonisation fronts (Ward 2007), particularly when using high confidence absence data.

While it is possible to estimate or model the probability of detection by repeated surveys and, hence, to discern observed true and false absences (MacKenzie and Royle 2005; Guillera-Arroita et al. 2015), it is often logistically impractical. Moreover, presences often include numerous opportunistic observations, whereas absences generally go unrecorded and suffer from greater uncertainty. Therefore, contrasting presences with background points, i.e. pseudo-absences, which are randomly sampled from within the study area (Iturbide et al. 2015), are the only viable option for many situations (Elith and Leathwick 2009). In this context, it has to be stressed that the default number of background points (10,000) in MAXENT, a broadly adopted maximum entropy machine-learning SDM algorithm (Phillips et al. 2004, 2017), is often insufficient and – unless increased sufficiently to capture the range of existing environmental conditions – equates to modelling at lower spatial resolution (Renner and Warton 2013).

When using presence-background occurrence data, the measured performance of modelling techniques, such as Generalised Linear Models (Nelder and Wedderburn 1972) or MAXENT (Phillips et al. 2004, 2017; Philipps and Dudík 2008), generally has a positive relationship with both the number and geographic extent of random background points relative to presence points (Philipps and Dudík 2008; Barbet-Massin et al. 2012). Model performance then, should be interpreted as the result of a complex interplay of artifacts (stemming from data or methods) and biological causes, since performance depends on modeller decisions, data availability and the underlying distribution of species based on dispersal from historical distributions (Barve et al. 2011). For instance, specialist species are, by definition, confined to narrower conditions within a broad landscape, than are generalist species. Unfortunately, the relative ease of characterising narrowly confined species (Jiménez-Valverde et al. 2008) and consequent high model performance scores, may not be due to biological causes, such as ecological specialisation, i.e. dependence on specific environmental conditions. Instead, good model performance may result from model overfitting as species presences simply coincide with specific conditions, combinations or transformations of environmental variables in overly complex models (Merow et al. 2014; Fourcade et al. 2018). Therefore, in order to interpret and compare the performance of models in a meaningful way, modellers must move past simple evaluation metrics (e.g. sensitivity and specificity) and consider confounding effects on model performance. Bias, cell number totals in the confusion matrix and prevalence should be assessed and their effect on model performance minimised by, for example, choosing less susceptible evaluation metrics.

The issue of confounding effects on model performance is particularly important in conservation planning and reserve area selection, since both regularly take SDM predictions into account (Margules and Pressey 2000; Lin et al. 2014; Guillera-Arroita et al. 2015). Unless confounding effects are considered during model evaluation, however, any application of SDM is potentially affected, including estimates of species richness and community composition (Gioia and Pigott 2000; Pineda and Lobo 2009;

Thuiller et al. 2015) or hindcasting past distributions (Franklin et al. 2015). Unfortunately, modellers often assume that model performance constitutes the most objective, if not the best, evidence for model legitimacy, representing not only prediction accuracy, but also underlying biological processes (Jiménez-Valverde et al. 2008). This assumption can be problematic as all evaluation metrics differentially react to modelling conditions. That is, two different evaluation metrics may represent true model performance better under varying circumstances (Lawson et al. 2014). More generally, evaluation metrics are functions of both the model prediction accuracy and the modelling conditions (Woodcock 1976). This means that beyond representing prediction accuracy, model evaluation metrics can conflate both artefactual (e.g. differences in sampling regimes, study extent, resolution, model overfitting) and biological (e.g. degree of species specialisation, population dynamics, autecology) signals.

In summary, the interpretability of measured model performance garnered from presence-background data is limited (Hirzel et al. 2006). The extent to which modelled predictions do reflect the posited goal of most initiatives – namely, identifying the underlying biological processes that dictate species distributions, is less certain. However, any measure of model performance for any given model is the result of a four-part process that includes data collection, model training, threshold setting and the selection of model evaluation metrics. Here, we focus on the fourth part only, the selection of appropriate model evaluation metrics under specific modelling conditions commonly encountered in presence-background SDM initiatives. We also provide some insights into how measured model performance may have resulted from biological signals or artifacts.

In this paper, we use purely theoretical scenarios to compare the responses of three evaluation metrics, the True Skill statistic (TSS; Allouche et al. 2006), the Odds Ratio Skill Score (ORSS; Stephenson 2000) and the Symmetric Extremal Dependence Index (SEDI; Ferro and Stephenson 2011), to three confounding factors on model performance, the cell number totals in the confusion matrix (typically dominated by background points total and dependent on the size of the study area and resolution), bias and prevalence. To our knowledge, the latter two evaluation metrics have not been used in SDM before. We also contextualise our results in terms of SDM initiatives with respect to how well specific evaluation metrics reflect biological signals versus artifacts in particular modelling conditions and discuss this with reference to recently raised concerns about the use of TSS in SDM reported in literature. Additionally, R code is provided for ORSS and SEDI computations.

Materials and methods

Comparison of evaluation metrics

Detailed definitions of some of the more technical terms and a comparison of the mathematical properties of the analysed evaluation metrics are found in Table 2.

Table 2. Comparison of selected properties of binary evaluation metrics compared in this article. 'Consistent at maxSSS' refers to the threshold maximising the sum of sensitivity and specificity (SSS) suggested by Liu et al. (2013) which is generally recommended in literature. Please refer to Somodi et al. (2017) for prevalence effects on maxSSS itself. "+" and "–" indicate that the evaluation metric features or lacks a property, respectively.

Property	Definition	TSS	ORSS	SEDI
Asymptotically equitable	Random predictions yield a score of zero	+	+	+
Prevalence independent	Same result when prevalence changes if both H and F remain unchanged	+	+	+
Complement symmetric	Same result when switching a and c with d and b	+	+	+
Consistent at maxSSS	Maximising SSS maximises the evaluation metric	+	–	–
Fixed range	Minimum/maximum possible values do not depend on prevalence	+	+	+
Hard to hedge	Monotonic increase with H and monotonic decrease with F	+	+	+
Non-degenerate	Meaningful results when prevalence approaches zero	–	–	+
Regular	Isopleths of the evaluation metric pass through the origin	–	+	+
Transpose symmetric	Same result when swapping b and c	–	+	–

Below are the equations of four simple evaluation metrics (variables a, b, c and d according to Table 1). First, the hit rate (H, also termed sensitivity) measures the ratio of true presences to the sum of true presences and omission errors while completely ignoring commission errors and the number of true absences. Second, the false positive rate (F), equal to 1–*specificity*, measures the ratio of commission errors to the sum of commission errors and true absences. Third, bias (also termed bias score or frequency bias) measures the ratio of commissions to omissions and helps to identify over-/under-predicting models. And fourth, prevalence (also termed base rate) measures the ratio of presences (both predicted and omitted) to all cells and hence expresses how common within the study area a species is, according to the available data.

$$H = a/(a + c) \tag{1}$$
$$F = b/(b + d) \tag{2}$$
$$bias = (a + b)/(a + c) \tag{3}$$
$$prevalence = (a + c)/(a + b + c + d) \tag{4}$$

TSS measures the difference between H and F and was first developed as Peirce's skill score in meteorology (Peirce 1884). It was later introduced to other fields, including the field of SDM, where it replaced kappa (Cohen 1960) and its strong unimodal response to prevalence (Allouche et al. 2006).

$$TSS = H - F \tag{5}$$

ORSS measures skill compared to a random prediction, is a synonym of Yule's Q (1900) and was introduced to meteorology by Stephenson (2000). ORSS provides equal error weighting but rapidly converges to one even for imperfect predictions

(Woodcock 1976) and hence, requires significance testing to quantify skill and to discern real skill from chance (Stephenson 2000).

$$ORSS = (ad - bc)/(ad + bc) \tag{6}$$

Ferro and Stephenson (2011) developed the Symmetric Extremal Dependence Index (SEDI) as an improvement on earlier work by Stephenson et al. (2008) and Hogan et al. (2009). SEDI featured greatly reduced sensitivity to prevalence, while retaining most beneficial properties of its predecessors (Ferro and Stephenson 2011). This includes asymptotic equitability, i.e. the ability to distinguish random and skilled predictions at smaller than infinite sample sizes (Hogan et al. 2010). Yet, SEDI is not applicable if any of the four cells in the confusion matrix equals zero (Ferro and Stephenson 2011) since *log*(0) yields infinity. Overfitted or misspecified models in these instances, however, can still be interpreted by adding an infinitely small number to those cells containing zeros. Our implementation of SEDI (see Supplementary Information for repository link) also issues a character string indicating if such approximations were used and how to best interpret the result.

$$SEDI = \frac{\log(F) - \log(H) - \log(1-F) + \log(1-H)}{\log(F) + \log(H) + \log(1-F) + \log(1-H)} \tag{7}$$

Theoretical scenarios

Using two types of extreme prediction settings ("Optimistic" and "Pessimistic") and two types of typical species prevalence settings ("Differential Bias" and "Changing Bias"), we investigated the response of TSS, ORSS and SEDI to increasing cell number total in the confusion matrix, varying commission error rates, omission error rates and species prevalence. These settings were each divided into two theoretical scenarios: "Incorrectly Optimistic" (IO) and "Correctly Optimistic" (CO), "Correctly Pessimistic" (CP) and "Incorrectly Pessimistic" (IP), "Commission Bias" (CB) and "Omission Bias" (OB) and "Low Commission Rate" (LC) and "High Commission Rate" (HC). For each of the eight scenarios (Table 3), we prepared twenty cases, i.e. confusion matrices, to be evaluated (Table A3). In "Changing Bias" scenarios (LC and HC), the term 'logistic' was used to describe model fit which improved with background point totals and cell number totals but plateaued below perfect fit.

Although not mutually exclusive, each scenario is designed to reflect signals that could have arisen from biological signals or artifacts, thereby revealing how susceptible model evaluation metrics are to conflating the two. Below, we briefly describe all scenarios and Fig. 1 visualises selected example cases. Across the theoretical cases in the different scenarios, the cell number total ($a + b + c + d$) approximately ranges from 1240 to 42560 with increasingly large step size (Table 3) to allow the analysis of model evaluation across a large range of modelling conditions without requiring an overly large number of cases. The R code to reproduce our analysis is available upon request.

Table 3. Description of the theoretical scenarios, where IO, CO, CP, IP, CB, OB, LC and HC are abbreviations for scenarios "Incorrectly Optimistic", "Correctly Optimistic", "Correctly Pessimistic", "Commission Bias", "Omission Bias", "Low Commission Rate" and "High Commission Rate". True positives, false positives, false negatives and true negatives are represented by a, b, c, and d, respectively. Total lists the sum of all four cells in the confusion matrix. The formulations are provided in pseudo-R-code, i.e. square brackets ("[" and "]") indicate vectors and colons (":") indicate a series. For example, "[x:y]" represents a vector of integers ranging from x to y. "..." are used to indicate repeating the same number, and n is the case number.

Scenario	a	b	c	d	Total
IO	increasing: $0.005*[1:n]^{1.25}*1000+248$	increasing: $[1:n]^{1.25}*1000+248-[a]$	constant: [1...1]	constant: [1...1]	min: 1250 max: 42545
CO	increasing: $0.995*[1:n]^{1.25}*1000+248$	increasing: $[1:n]^{1.25}*1000+248-[a]$	constant: [1...1]	constant: [1...1]	min: 1250 max: 42545
CP	constant: [1...1]	constant: [1...1]	increasing: $[1:n]^{1.25}*1000+248-[a]$	increasing: $0.995*[1:n]^{1.25}*1000+248$	min: 1250 max: 42545
IP	constant: [1...1]	constant: [1...1]	increasing: $[1:n]^{1.25}*1000+248-[a]$	increasing: $0.005*[1:n]^{1.25}*1000+248$	min: 1250 max: 42545
CB	constant: [200...200]	constant: [30...30]	constant: [20...20]	increasing: $[1:n]^{1.25}*1000$	min: 1250 max: 42545
OB	constant: [200...200]	constant: [20...20]	constant: [30...30]	increasing: $[1:n]^{1.25}*1000$	min: 1250 max: 42545
LC	'logistic': [175:189,190...190]	increasing: $[c_n:c_1]$	decreasing: 200-[a]	increasing: $[1:n]^{1.25}*1000$	min: 1210 max: 42520
HC	'logistic': [175:189,190...190]	increasing: $3*[c_n:c_1]$	decreasing: 200-[a]	increasing: $[1:n]^{1.25}*1000$	min: 1230 max: 42570

Scenarios IO, CO, CP and IP were designed to demonstrate how evaluation metrics at essentially constant extreme levels of prevalence react to an increasing cell number total in the confusion matrix. The biological component of these scenarios is analogous to specialist or generalist species that have a constant prevalence of 0.5% or 99.5% of the study area. The artefactual component is related to the implications of study area increases for the number of background points and total number of cells and their effect on the calculation of evaluation metrics. Scenario IO was characterised by large numbers of commission errors as it evaluated an extreme incorrectly optimistic modelling prediction (over-prediction) when true species prevalence is equal to 0.5%, reflecting extreme specialisation or rarity and under increasing background size. Scenario CO was identical to scenario IO in its extreme prediction. However, as true species prevalence was equal to 99.5% (reflecting extremely low specialisation), it no longer resembled an over-prediction and was consequently dominated by true presences. Scenario CP evaluated an extreme correctly pessimistic prediction when true species prevalence was equal to 0.5%, reflecting a high degree of ecological specialisation and species presence was only predicted for a small proportion of the study area, under increasing background size. This scenario was characterised by large numbers of true absences. Scenario IP was identical to scenario CP in its extreme prediction but dominated by false

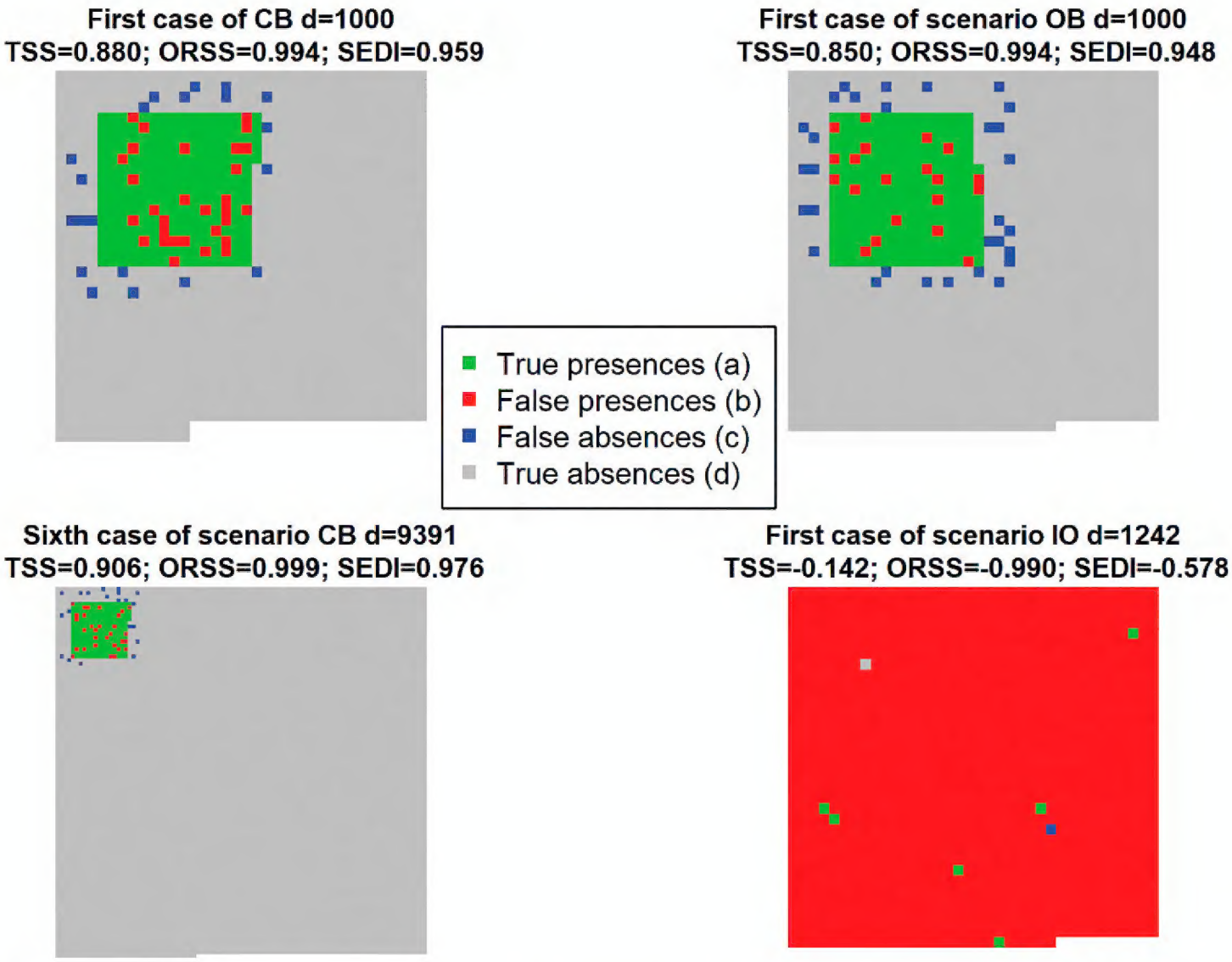

Figure 1. Potential spatial distributions of confusion matrix categories corresponding to values of the True Skill Statistic (TSS), the Odds Ratio Skill Score (ORSS) and the Symmetric Extremal Dependence Score (SEDI) for selected scenario cases.

absences since true species prevalence was now equal to 99.5%, turning it into a gross under-prediction.

Scenarios CB and OB were designed to reveal the effect of bias on evaluation metrics under decreasing prevalence (~17% to ~0.5%) as the study area increased. In these scenarios, evaluation metrics should consistently penalise model predictions according to the degree of their bias, across the whole range of prevalence. Scenario CB was more optimistic (more commission errors and more predicted presences) than scenario OB (more omissions and fewer predicted presences). Therefore, the two scenarios together can be seen as a test of transpose symmetry.

Scenarios LC and HC examined the response of evaluation metrics to changes in bias while model fit (i.e. the number of true positives) and the total number of cells increased as prevalence decreased (~17% to ~0.5%). More specifically, the number of observations was held constant in both scenarios, while the numbers of true positives and omissions increased and decreased, respectively. However, at the same time, commission errors became more frequent. In other words, the bias of the model changed together with prevalence and the size of the study area. Scenarios LC and HC differed only in their rate of commission errors which was three times higher in scenario HC

Table 4. Evaluation scores (rounded to four digits) for all evaluation measures metrics considered across all scenarios. IO, CO, CP, IP, CB, OB, LC and HC are abbreviations for scenarios "Incorrectly Optimistic", "Correctly Optimistic", "Correctly Pessimistic", "Commission Bias", "Omission Bias", "Low Commission Rate" and "High Commission Rate". Total lists the sum of all four cells in the confusion matrix. H, F, TSS, ORSS and SEDI list evaluation metric values for hit rate, false positive rate, True Skill Statistic, Odds Ratio Skill Score and Symmetric Extremal Dependence Score, respectively. Cases #6 and #7 closely resemble typical presence-background modelling conditions in MAXENT.

Cell Total	Scenario	H	F	TSS	ORSS	SEDI
ca. 9,000 – 10,000 (Case #6)	IO	0.9796	0.9999	-0.0203	-0.9900	-0.4050
	CO	0.9999	0.9796	0.020	0.9900	0.4050
	CP	0.0204	0.0001	0.0203	0.9900	0.4050
	IP	0.0001	0.0204	-0.0203	-0.9900	-0.4050
	CB	0.9091	0.0032	0.9059	0.9994	0.9761
	OB	0.8696	0.0021	0.8674	0.9994	0.9659
	LC	0.9000	0.0012	0.8988	0.9997	0.9767
	HC	0.9000	0.0035	0.8965	0.9992	0.9730
ca. 11,000 – 12,000 (Case #7)	IO	0.9831	0.9999	-0.0169	-0.9900	-0.3937
	CO	0.9999	0.9831	0.0169	0.9900	0.3937
	CP	0.0169	0.0001	0.0169	0.9900	0.3937
	IP	0.0001	0.0169	-0.0169	-0.9900	-0.3937
	CB	0.9091	0.0026	0.9065	0.9995	0.9768
	OB	0.8696	0.0018	0.8678	0.9995	0.9668
	LC	0.9050	0.0011	0.9039	0.9998	0.9783
	HC	0.9050	0.0032	0.9018	0.9993	0.9749

than in scenario LC. The biological component could represent increasing specialisation of a given species as the study extent increases; whereas the artefactual component could represent resultant increases in model fit as increasing specialisation makes for easier characterisation (Jiménez-Valverde et al. 2008).

Results

Our results are summarised in Table 4 and presented in Fig. 2. In addition, we provide Fig. 3 which depicts the proportional difference between scores of SEDI and TSS, ORSS and TSS and SEDI and ORSS, for scenarios LC and HC, i.e. the sensitivity to changes in commission errors under decreasing prevalence.

Optimistic and pessimistic scenarios

TSS shows a strong response to increased study area size and, hence, confusion matrix cell number totals and rapidly converges to zero, rendering indifferent useful and random models beyond cell number totals in the confusion matrix of approximately

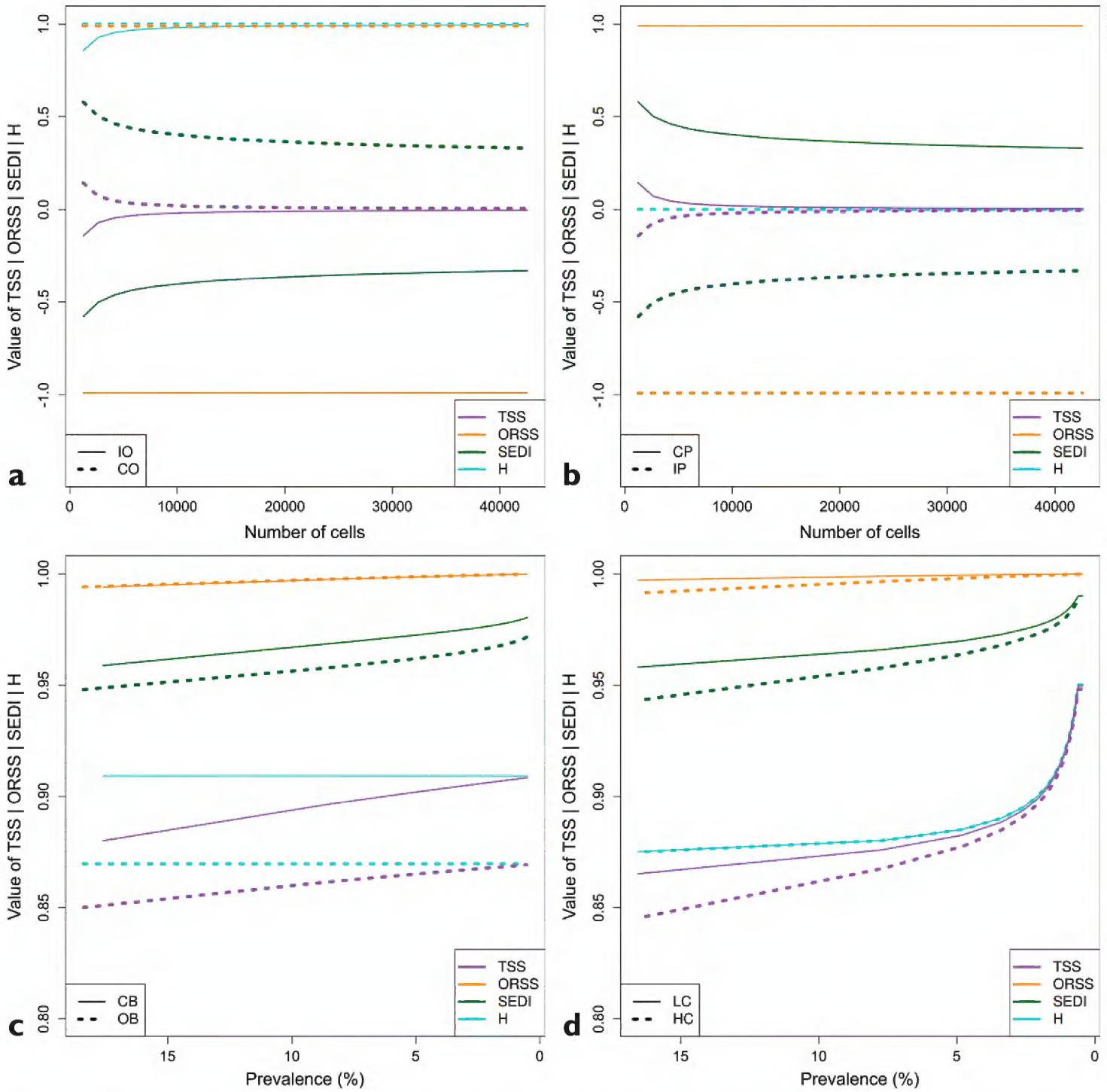

Figure 2. Plots of the values of hit rate (H), the True Skill Statistic (TSS), the Odds Ratio Skill Score (ORSS) and the Symmetric Extremal Dependence Score (SEDI) for all eight scenarios. Panels **a, b, c, d** display scenarios "Incorrectly Optimistic" and "Correctly Optimistic", "Correctly Pessimistic" and "Incorrectly Pessimistic", "Commission Bias" and "Omission Bias" and "Low Commission Rate" and "High Commission Rate". In panels a and b, the x-axis denotes the log of the total number of cells, i.e. the size of the study area, whereas in panels c and d, the x-axis denotes prevalence (%).

30,000 cells. SEDI shows only a moderate response and converges much later to zero. Of note, H completely fails in this respect since both incorrectly optimistic predictions (IO) and correctly pessimistic predictions (CP) converge to one and zero, respectively and yield scores very similar to those of their correct (CO) and incorrect (IP) counterparts. Finally, SEDI has stronger discriminatory power than TSS at intermediate study areas yet, only ORSS is expected to correctly assess model performance as study area size converges to infinity (Fig. 2a, b).

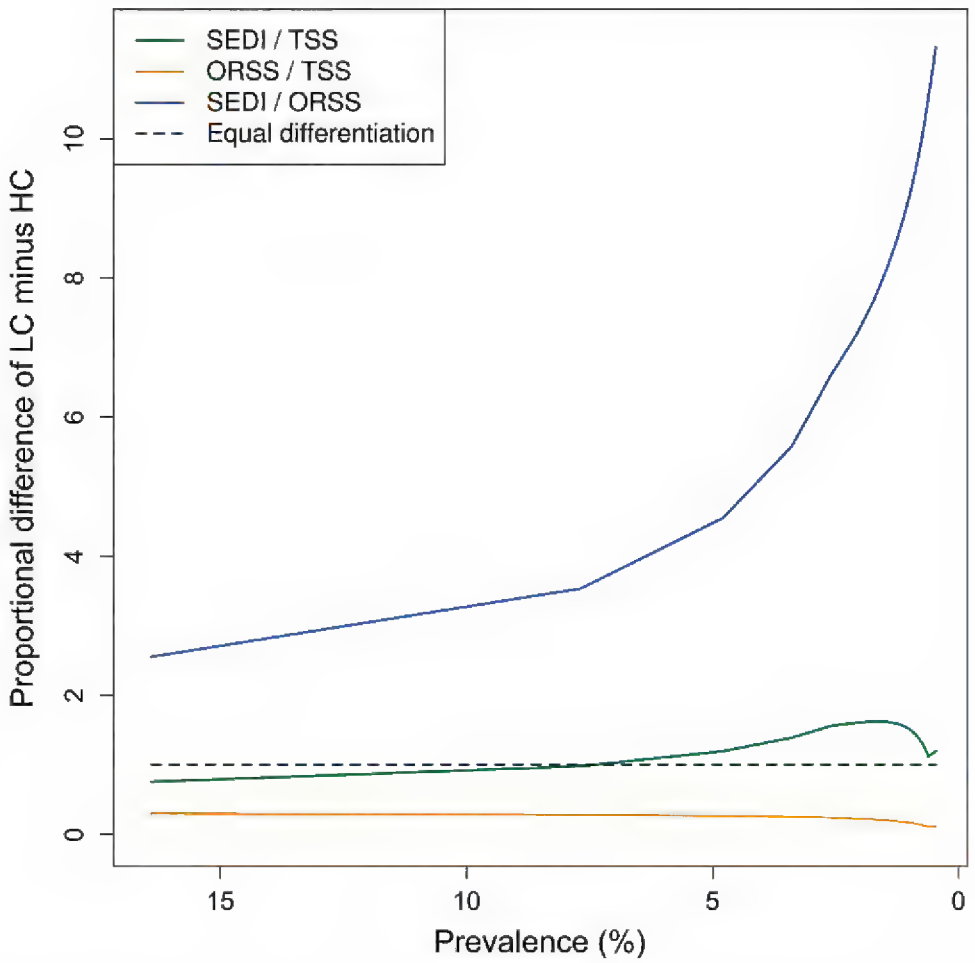

Figure 3. Proportional difference between scenarios "Low Commission Rate" (LC) and "High Commission Rate" (HC) for (in dark green) the Symmetric Extremal Dependence Score (SEDI) vs. the True Skill Statistic (TSS), (in orange) the Odds Ratio Skill Score (ORSS) vs. TSS and (in blue) SEDI vs. ORSS. The black, horizontal, dashed line represents equal differentiation. As there are slight differences in prevalence between scenarios LC and HC, the x-axis shows the mean prevalence for given cases across both scenarios.

Differential bias scenarios and changing bias scenarios

TSS quickly converges with H and always favours over-predictions to under-predictions. However, the degree to how much over-predictions are favoured increases as prevalence decreases. Although SEDI also favours over-predictions, it does so to a much smaller degree and is not significantly affected by prevalence. Just as in scenarios CO and CP, ORSS rapidly converges to one (Fig. 2c). Under increasing study area and background point totals and so decreasing prevalence, TSS converges quickly with H for the most realistic scenarios (LC and HC in Fig. 2d) and models that predict increasing amounts of both true presences and commission errors, as omission errors, decrease. At higher levels of prevalence, TSS can still discern the quality of models differing only in their rate of commission errors, but once prevalence falls below approximately 2.5%, their difference becomes indistinguishable. SEDI can assess the quality of models differing only in their rate of commission errors as prevalence decreases to almost zero. As in previous scenarios, since ORSS rapidly converges to one, model scores differing only in their rate of commission errors become indistinguishable even faster than when using TSS. In addition, the proportional difference between LC and HC SEDI scores and TSS scores are lower at the start though increase as prevalence

decreases (Fig. 3). This indicates that, although TSS identifies differences in commission error levels at high prevalence (greater than approximately 7%) better than SEDI, the reverse is true at the low prevalence levels typically encountered in presence-background modelling.

Discussion

Using eight theoretical scenarios, we have shown that TSS, ORSS and SEDI, as well as their underlying evaluation measures (H and F, see F in Table 2), show distinct responses to: 1) increasing size of the study area and, hence, growing numbers of background points, even when prevalence is kept constant (scenarios IO, CO, CP and IP), 2) to the direction of bias as prevalence decreases and the extent of the study area and cell number totals increase (scenarios CB and OB) and 3) to changes in bias as prevalence decreases and the extent of the study area and cell number totals increase (scenarios LC and HC).

Our analysis confirmed a very problematic property of TSS. That is, a very large number found in any of the four cells of the confusion matrix (Table 1) leads to the marginalisation of the other entry in the same column (Stephenson 2000). This means that, when assessing rare events, such as rare species presence, TSS quickly converges to H (Doswell et al. 1990). Less apparent responses of TSS to prevalence have also been discussed in the field of SDM, for instance, by Somodi et al. (2017) who found that small sample size exacerbates the effects of prevalence on TSS. We also evaluated two alternative evaluation metrics from the field of meteorology, SEDI and ORSS. The former appears to be ideal for typical low prevalence presence-background SDM conditions, whereas the latter may be useful for high confidence presence-absence data or if strictly equal error weighting is required.

Optimistic and pessimistic scenarios

By grossly over- or under-predicting the distribution of a hypothetical target species, we observed the response of evaluation metrics to extreme biases in less realistic scenarios. These extreme scenarios, however, have also shown that discernment of strongly and weakly performing models greatly differs amongst evaluation metrics and modelling conditions. While these scenario results support the use of ORSS for large study extents, because of its rapid convergence to one, even for imperfect predictions (Woodcock 1976), significance testing is required in order to determine the quality of and to allow comparisons across models (Stephenson 2000). Therefore, SEDI is preferable as it allows direct assessments of model quality across a much larger spectrum of study extents. Further, SEDI assessments are made with higher discriminatory power than TSS, which rapidly converges to zero for extreme predictions.

Differential bias and changing bias scenarios

These scenarios have been designed to reflect common modelling conditions in order to observe the response of evaluation measures to differential (CB and OB) and changing (LC and HC) biases, under decreasing prevalence as the size of the study area and, hence, the number of background points increased. Analysis of these scenarios revealed very distinct responses to the differing modelling conditions. Results for scenarios CB and OB and LC and HC suggest the use of SEDI since: 1) TSS encourages overpredictions due to its strongly biased treatment of errors which increases as prevalence decreases; 2) TSS quickly loses the discriminatory power to differentiate between models, differing only in their commission rate as it always converges to H; and 3) ORSS converges to one so rapidly (Stephenson 2000) that such differences vanish at even higher prevalence levels than when using TSS.

Modelling conditions and research questions

Our analysis reaffirms the importance of selecting model evaluation metrics corresponding with modelling questions and conditions (Woodcock 1976). Important modelling conditions include the extent of the study area (Termansen et al. 2006), prevalence (Doswell et al. 1990; Stephenson 2000; Somodi et al. 2017; Leroy et al. 2018) and the relative severity of model error types due to varying degrees of biological and artefactual causes. Important biological factors, related to the extent of the study area and species prevalence, include the degree of specialisation within the given landscape (Jiménez-Valverde et al. 2008); population sinks (Guisan and Thuiller 2005); and species equilibrium (Václavík and Meentemeyer 2012). This latter factor is also important in the context of invasive species (Ward 2007). Artifacts can originate from the cell number total in the confusion matrix which is largely driven by background points and hence dependent on modelling resolution (Seo et al. 2008) and study extent. Artefactual signals can also be caused by prevalence and its interactions with sample-size (Somodi et al. 2017) and the degree of confidence in absence data (Leroy et al. 2018) which is influenced by species mobility (Jaberg and Guisan 2001) and a multitude of other factors.

Do commission errors matter?

Our results suggested a limited capacity of TSS to provide consistent performance comparisons across varying modelling conditions. This is worrying because TSS may yield misleading estimates of model fidelity, which can lead to the selection of inadequate models. Although it may be tempting to assume that researchers would recognise anomalous conditions where TSS scores are misleading (such as those presented here), this is not necessarily the case – as demonstrated by the broad and seemingly uncritical applica-

tion of TSS in presence-background SDM over the last decade. Complications, owing to the relative inability of TSS to provide information on commission errors as prevalence approaches zero, are more nuanced. Ultimately, such complications are only problematic in as much as commission errors matter, which depends on *bias* > 1, the question, the available data and the biology of the species. In fact, it is widely acknowledged that even absence data from professional surveys have greater degrees of both sampling and ecological uncertainty than presence data (MacKenzie et al. 2017). Therefore, one could argue that errors of omission, i.e. predicting absences where there are species observations, are far more grievous than commission errors in model evaluation, particularly when using presence-background data. This is because commission errors can be relatively obscure and should be given less weight than omission errors in model evaluation (Braunisch and Suchant 2010; Liu et al. 2013). Taking the above reasoning to its logical extreme then, commission errors are irrelevant when considering pseudo-absence data.

While the above reasoning is persuasive, simply ignoring commission errors in presence background data by limiting evaluation to H, is not a viable option under all but a small subset of questions, modelling conditions and biological assumptions. More specifically, doing so would be incongruent with the biological circumstances, sampling realities and the intents of most modelling initiatives. Further, evaluation scores would become more vulnerable to artificial inflation. From a biological perspective, model evaluation metrics that ignore commission errors are equivalent to assuming that all background points are locations where the species is present but unobserved. That is, assuming that observed presence locations may represent the subset of relative high use or occupied conditions within local settings (Elith et al. 2011). While this may be a good assumption for situations where habitat use of a generalist species is considered (particularly if survey effort has been uniform across the entire study area), it is unrealistic and impractical for many modelling initiatives that attempt to characterise either the fundamental or realised niche (Elith and Leathwick 2009), based on incomplete sampling regimes.

Furthermore, even when the above biological assumptions and survey prerequisites are valid, explicitly choosing to ignore commission errors further assumes that unobserved locations are irrelevant—an assumption that is seldom the case since these locations may correspond to low population density areas (Guisan and Thuiller 2005). Models (and subsequent thresholds) that commit lower numbers of commission errors may better differentiate between high and low population density areas, a highly desirable characteristic for assessing population status (Tôrres et al. 2012). Of course in rarer situations, strictly considering omission errors in model evaluation may be desirable such as when assessing the impacts of decisions on vulnerable species (Karl et al. 2000). In addition, ignoring commission errors can lead to unintended consequences as seen in Fig. 2a where H is deceptively high for incorrectly optimistic outputs, i.e. over-predicting models. Although commission errors should be weighted less than omission errors in most SDM initiatives, as accomplished by both TSS and SEDI, this does not mean that they are irrelevant or become irrelevant when prevalence decreases. Exceptions to this are very specific circumstances.

This study demonstrated that more consistent commission error weighting (as with SEDI) also circumvents a number of potentially artefactual signals as prevalence

approaches zero. We also discussed the relative inability of TSS to compare performance across modelling conditions. For these reasons, whereas maximising TSS may be instrumental when presence-absence thresholds are required (Liu et al. 2013), TSS may not perform well in presence-background model evaluations (Somodi et al. 2017; Leroy et al. 2018). Unless circumstances require down-weighting of commission errors as prevalence decreases, SEDI's ability to take into account information on both error types across a wide range of modelling conditions makes it a better choice for low prevalence conditions, characteristic of presence-background modelling approaches.

The use of similarity measures as an alternative to TSS has recently been suggested by Leroy et al. (2018). However, similarity measures are only applicable when there is a known truth such as when modelling virtual species (see Hirzel et al. 2001; Leroy et al. 2016 for an introduction to virtual species). One might also consider the use of bootstrapping and related techniques (Efron 1983) such as sub-sampling in model evaluation (Verbyla and Litvaitis 1989), i.e. the repeated sub-sampling of background points to the number of presences and averaging of the resulting scores. Sub-sampling would remove the bias caused by large numbers of background points towards true absences from confusion matrices altogether, albeit at the cost of being computationally intensive. Therefore, alternative measures such as SEDI or, for some specific cases, ORSS, appear superior as they are neither limited to virtual species nor costly in terms of computation.

Conclusions

In our study, we focused on the importance of model evaluation in the context of ecology and conservation. The problems discussed are particularly relevant in systematic conservation planning (Margules and Pressey 2000; Lin et al. 2014; Guillera-Arroita et al. 2015) but will likely cause issues in any application of SDM. This is, since in any application, different scores favour different models (e.g. over-predictive vs. under-predictive, as they differentially respond to modelling conditions (Woodcock 1976), which inevitably affects outcomes.

Our results indicate that ORSS is a suitable evaluation metric for high-confidence presence-absence data, high prevalence situations or if strictly equal error weighting is required. SEDI and to a lesser degree TSS, are suitable evaluation metrics for presence-background SDM initiatives, since the error weighting of the evaluation metrics better reflects low-confidence pseudo-absence data. However, since SEDI provides more consistent performance scores and weighting of commission errors over a wide range of study extents (and background point totals) and prevalence, it is better suited for presence-background SDM, which is applied over a wide range of modelling conditions (i.e. to common or rare species and across single protected areas or whole continents). Finally, we strongly recommend abstaining from the use of TSS whenever prevalence is lower than approximately 2.5% or when a large number of background points is used that drives the total number of cells in the confusion matrix to more than roughly 30,000 cells since TSS will not distinguish between low and high commission error rates or useful and random models.

Acknowledgements

We are grateful to our reviewers Eduardo Arle and Paul Holloway and Petr Keil for their helpful comments and suggestions. We also thank P.A. Château for his comments on an earlier version of this manuscript.

References

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43(6): 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3(2): 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, Soberón J, Villalobos F (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling 222(11): 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Bascompte J (2009) Mutualistic networks. Frontiers in Ecology and the Environment 7(8): 429–436. https://doi.org/10.1890/080026

Braunisch V, Suchant R (2010) Predicting species distributions based on incomplete survey data: The trade-off between precision and scale. Ecography 33(5): 826–840. https://doi.org/10.1111/j.1600-0587.2009.05891.x

Bulleri F, Bruno JF, Silliman BR, Stachowicz JJ (2016) Facilitation and the niche: Implications for coexistence, range shifts and ecosystem functioning. Functional Ecology 30(1): 70–78. https://doi.org/10.1111/1365-2435.12528

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1): 37–46. https://doi.org/10.1177/001316446002000104

Doswell CA III, Davies-Jones R, Keller DL (1990) On summary measures of skill in rare event forecasting based on contingency tables. Weather and Forecasting 5(4): 576–585. https://doi.org/10.1175/1520-0434(1990)0052.0.CO;2

Efron B (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association 78(382): 316–331. https://doi.org/10.1080/01621459.1983.10477973

Elith J, Leathwick JR (2009) Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology Evolution and Systematics 40(1): 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Diversity & Distributions 17(1): 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

Ferro CAT, Stephenson DB (2011) Extremal dependence indices: Improved evaluation measures for deterministic forecasts of rare binary events. Weather and Forecasting 26(5): 699–713. https://doi.org/10.1175/WAF-D-10-05030.1

Fordham DA, Akçakaya HR, Araújo MB, Brook BW (2012) Modelling range shifts for invasive vertebrates in response to climate change. In: Brodie J, Post E, Doak D (Eds) Wildlife conservation in a changing climate. The University of Chicago Press, Chicago, 86–108.

Fourcade Y, Besnard AG, Secondi J (2018) Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Global Ecology and Biogeography 27(2): 245–256. https://doi.org/10.1111/geb.12684

Franklin J, Potts AJ, Fisher EC, Cowling RM, Marean CW (2015) Paleodistribution modeling in archaeology and paleoanthropology. Quaternary Science Reviews 110: 1–4. https://doi.org/10.1016/j.quascirev.2014.12.015

Gaston K (1994) Rarity. Chapman & Hall (London): 1–205. https://doi.org/10.1007/978-94-011-0701-3

Gioia P, Pigott JP (2000) Biodiversity assessment: A case study in predicting richness from the potential distributions of plant species in the forests of south-western Australia. Journal of Biogeography 27(5): 1065–1078. https://doi.org/10.1046/j.1365-2699.2000.00461.x

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135(2–3): 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Guisan A, Thuiller W (2005) Predicting species distribution: Offering more than simple habitat models. Ecology Letters 8(9): 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA (2015) Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24(3): 276–292. https://doi.org/10.1111/geb.12268

Hanski I (1998) Metapopulation dynamics. Nature 396(6706): 41–49. https://doi.org/10.1038/23876

Hardin G (1960) The competitive exclusion principle. Science 131(3409): 1292–1297. https://doi.org/10.1126/science.131.3409.1292

Heilbron DC (1994) Zero-altered and other regression models for count data with added zeros. Biometrical Journal. Biometrische Zeitschrift 36(5): 531–547. https://doi.org/10.1002/bimj.4710360505

Hirzel AH, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species. Ecological Modelling 145(2–3): 111–121. https://doi.org/10.1016/S0304-3800(01)00396-9

Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199(2): 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hogan RJ, O'Connor EJ, Illingworth AJ (2009) Verification of cloud-fraction forecasts. Quarterly Journal of the Royal Meteorological Society 135(643): 1494–1511. https://doi.org/10.1002/qj.481

Hogan RJ, Ferro CA, Jolliffe IT, Stephenson DB (2010) Equitability revisited: Why the "equitable threat score" is not equitable. Weather and Forecasting 25(2): 710–726. https://doi.org/10.1175/2009WAF2222350.1

Iturbide M, Bedia J, Herrera S, del Hierro O, Pinto M, Gutiérrez JM (2015) A framework for species distribution modelling with improved pseudo-absence generation. Ecological Modelling 312: 166–174. https://doi.org/10.1016/j.ecolmodel.2015.05.018

Jaberg C, Guisan A (2001) Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. Journal of Applied Ecology 38(6): 1169–1181. https://doi.org/10.1046/j.0021-8901.2001.00668.x

Jiménez-Valverde A, Lobo JM, Hortal J (2008) Not as good as they seem: The importance of concepts in species distribution modelling. Diversity & Distributions 14(6): 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x

Karl JW, Heglund PJ, Garton EO, Scott JM, Wright NM, Hutto RL (2000) Sensitivity of species habitat-relationship model performance to factors of scale. Ecological Applications 10(6): 1690–1705. https://doi.org/10.1890/1051-0761(2000)010[1690:SOSHRM]2.0.CO;2

Lawson CR, Hodgson JA, Wilson RJ, Richards SA (2014) Prevalence, thresholds and the performance of presence-absence models. Methods in Ecology and Evolution 5(1): 54–64. https://doi.org/10.1111/2041-210X.12123

Leathwick JR, Austin MP (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology 82(9): 2560–2573. https://doi.org/10.1890/0012-9658(2001)082[2560:CIBTSI]2.0.CO;2

Leroy B, Meynard CN, Bellard C, Courchamp F (2016) virtualspecies, an R package to generate virtual species distributions. Ecography 39(6): 599–607. https://doi.org/10.1111/ecog.01388

Leroy B, Delsol R, Hugueny B, Meynard CN, Barhoumi C, Bellard C (2018) Without quality presence-absence data, discrimination metrics such as TSS can be misleading measures of model performance. Journal of Biogeography 45(9): 1994–2002. https://doi.org/10.1111/jbi.13402

Lin YP, Huang CW, Ding TS, Wang YC, Hsiao WT, Crossman ND, Lengyel S, Lin WC, Schmeller DS (2014) Conservation planning to zone protected areas under optimal landscape management for bird conservation. Environmental Modelling & Software 60: 121–133. https://doi.org/10.1016/j.envsoft.2014.06.009

Liu C, White M, Newell G (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of Biogeography 40(4): 778–789. https://doi.org/10.1111/jbi.12058

MacKenzie DI, Royle JA (2005) Designing occupancy studies: General advice and allocating survey effort. Journal of Applied Ecology 42(6): 1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83(8): 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey L, Hines JE (2017) Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence. 2nd ed., Academic Press (London): 1–648.

Margules CR, Pressey RL (2000) Systematic conservation planning. Nature 405(6783): 243–253. https://doi.org/10.1038/35012251

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. Ecology Letters 8(11): 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

Merow C, Smith MJ, Edwards Jr TC, Guisan A, McMahon SM, Normand S, Thuiller W, Wüest RO, Zimmermann NE, Elith J (2014) What do we gain from simplicity versus complexity in species distribution models? Ecography 37(12): 1267–1281. https://doi.org/10.1111/ecog.00845

Nelder JA, Wedderburn RWM (1972) Generalized linear models. Journal of the Royal Statistical Society A 135(3): 370–384. https://doi.org/10.2307/2344614

Peirce CS (1884) The numerical measure of the success of predictions. Science 4(93): 453–454. https://doi.org/10.1126/science.ns-4.93.453-a

Peterson AT (2001) Predicting species' geographic distributions based on ecological niche modeling. The Condor 103(3): 599–605. https://doi.org/10.1650/0010-5422(2001)103[0599:PSGDBO]2.0.CO;2

Phillips SJ, Dudík M, Schapire RE (2004) July. A maximum entropy approach to species distribution modeling. In Carla Brodely (Ed.) Proceedings of the twenty-first international conference on Machine learning, Banff (Canada), July 2004, Association for Computing Machinery (New York): 655–662.

Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 31(2): 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME (2017) Opening the black box: An open-source release of Maxent. Ecography 40(7): 887–893. https://doi.org/10.1111/ecog.03049

Pineda E, Lobo JM (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. Journal of Animal Ecology 78(1): 182–190. https://doi.org/10.1111/j.1365-2656.2008.01471.x

Renner IW, Warton DI (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69(1): 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Seo C, Thorne JH, Hannah L, Thuiller W (2008) Scale effects in species distribution models: Implications for conservation planning under climate change. Biology Letters 5(1): 39–43. https://doi.org/10.1098/rsbl.2008.0476

Somodi I, Lepesi N, Botta-Dukát Z (2017) Prevalence dependence in model goodness measures with special emphasis on true skill statistics. Ecology and Evolution 7(3): 863–872. https://doi.org/10.1002/ece3.2654

Stephenson DB (2000) Use of the "odds ratio" for diagnosing forecast skill. Weather and Forecasting 15(2): 221–232. https://doi.org/10.1175/1520-0434(2000)0152.0.CO;2

Stephenson DB, Casati B, Ferro CAT, Wilson CA (2008) The extreme dependency score: A non-vanishing measure for forecasts of rare events. Meteorological Applications 15(1): 41–50. https://doi.org/10.1002/met.53

Termansen M, McClean CJ, Preston CD (2006) The use of genetic algorithms and Bayesian classification to model species distributions. Ecological Modelling 192(3–4): 410–424. https://doi.org/10.1016/j.ecolmodel.2005.07.009

Thuiller W, Pollock LJ, Gueguen M, Münkemüller T (2015) From species distributions to meta-communities. Ecology Letters 18(12): 1321–1328. https://doi.org/10.1111/ele.12526

Tôrres NM, De Marco P, Santos T, Silveira L, de Almeida Jácomo AT, Diniz-Filho JA (2012) Can species distribution modelling provide estimates of population densities? A case study with jaguars in the Neotropics. Diversity & Distributions 18(6): 615–627. https://doi.org/10.1111/j.1472-4642.2012.00892.x

Tyre AJ, Tenhumberg B, Field SA, Niejalke D, Parris K, Possingham HP (2003) Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. Ecological Applications 13(6): 1790–1801. https://doi.org/10.1890/02-5078

Václavík T, Meentemeyer RK (2012) Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Diversity & Distributions 18(1): 73–83. https://doi.org/10.1111/j.1472-4642.2011.00854.x

Verbyla DL, Litvaitis JA (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. Environmental Management 13(6): 783–787. https://doi.org/10.1007/BF01868317

Ward DF (2007) Modelling the potential geographic distribution of invasive ant species in New Zealand. Biological Invasions 9(6): 723–735. https://doi.org/10.1007/s10530-006-9072-y

Willis SG, Hill JK, Thomas CD, Roy DB, Fox R, Blakeley DS, Huntley B (2009) Assisted colonization in a changing climate: A test-study using two UK butterflies. Conservation Letters 2(1): 46–52. https://doi.org/10.1111/j.1755-263X.2008.00043.x

Wisz MS, Pottier J, Kissling WD, Pellissier L, Lenoir J, Damgaard CF, Dormann CF, Forchhammer MC, Grytnes JA, Guisan A, Heikkinen RK, Høye TT, Kühn I, Luoto M, Maiorano L, Nilsson M-C, Normand S, Öckinger E, Schmidt NM, Termansen M, Timmermann A, Wardle DA, Aastrup P, Svenning J-C (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. Biological Reviews of the Cambridge Philosophical Society 88(1): 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Woodcock F (1976) The evaluation of yes/no forecasts for scientific and administrative purposes. Monthly Weather Review 104(10): 1209–1214. https://doi.org/10.1175/1520-0493(1976)1042.0.CO;2

Yule GU (1900) Notes on the theory of association of attributes in statistics: With illustrations from the material of the childhood society, &c. Philosophical Transactions of the Royal Society A 194(252–261): 257–319. https://doi.org/10.1098/rsta.1900.0019